The positive role of the ecological community in the genomic revolution

Dawn Field, Barbara Methe, Karen Nelson, and Nikos Kyrpides,

An increasing number of ecologists are turning to the use of genomic technologies to complement and extend traditional methods of characterizing microbial diversity and its biological consequences (Xu, 2006). The SCOPE meeting on Microbial Environmental Genomics (MicroEnGen-II, Shanghai, June 12-15, 2006) provided a rich set of examples of progress at this interface (van Veen, Kowalchuk, 2006), many of which are described in this special issue.

The session of this meeting entitled 'new frontiers' included two presentations on informatics and data management. We (DF and NK) talked about projects we are involved in, namely, the Integrated Microbial Genomes Resource (IMG) (Markowitz *et al.*, 2006), the Genomic Standards Consortium (GSC) (Field *et al.*, et al, in review) and the NERC Environmental Bioinformatics Centre (NEBC) (Field *et al.*, 2005). In this commentary we briefly describe these projects, provide updated information on their development and attempt to expand on the wider issue of the increasing need for ecologists to become involved in the development of these and similar genomic resources.

These projects have all benefited from the direct and indirect contribution of ecologists. Further, each is working in its own way to help strengthen the position of researchers working in eco- and environmental genomics. The IMG system is an example of a new generation database project attempting to standardize access to the publicly available collection of genomes, through a comparative genomics approach (Markowitz *et al.*, 2006). The GSC represents an international effort to produce

consensus on the ways in which we collect, exchange and represent genomic metadata (Field *et al.*, et al, in review). NEBC is an example of a bioinformatics initiative in the United Kingdom designed to promote the uptake of 'omic technologies (genomics, transcriptomics, proteomics, metabolomics, etc) within the environmental genomics community (Field *et al.*, 2005).

The main message of this commentary is that ecologists should continue to play an active role, not only in the selection and analysis of samples for sequencing, but also, increasingly, in the generation and management of genomic resources (like databases, standards, and tools). Such involvement will not only help to make sure ecologists can exploit these resources but will also improve their value for the entire scientific community (Martiny and Field, 2005).

The growing availability of DNA from ecologically and environmentally important organisms

Genomic research has long been highly biased towards the characterization of model organisms and pathogens, but this imbalance is rapidly shifting (Martiny and Field, 2005; Nelson, 2000). Examples of large-scale projects in this area include the Gordon and Betty Moore Foundation's (GBMF) funding to sequence the genomes of 130 marine prokaryotes (http://www.moore.org/microgenome/) and the 'whole community' sequencing of samples from the J. Craig Venter Institute's Sorcerer II Expedition (http://www.sorcerer2expedition.org/) and the Community Sequencing Program from the Joint Genome Institute (http://www.jgi.doe.gov/CSP/index.html).

This shift may be just the tip of the iceberg as new, cheaper methods of sequencing come on board. A third presentation in the "new frontiers" session of this meeting (van Veen, Kowalchuk, 2006)

highlighted how ultra-high-throughput and low cost sequencing methods represent a quantum leap forward in our ability to acquire data at the DNA level (Shendure *et al.*, 2005). Specifically, Kun Zhang described work on the development of "Polony" sequencing, a process which can be used to determine a complete genome sequence from a single cell (Shendure *et al.*, 2005; Zhang *et al.*, 2006)

Polony is one of a family of new sequencing technologies (Shendure *et al.*, 2004) and pyrosequencing also holds great promise for the examination of environmental DNA (Clarke, 2005). It has, for example, been used to study metabolically different communities of microbes (Edwards *et al.*, 2006) and to obtain paleo-DNA from Mammoth (Poinar *et al.*, 2006). The potential outputs of these approaches are astonishing – the result of a single pyrosequencing run can produce up to 250,000 sequences, or 20 million base pairs. Applied to a single species, this can mean rapid completion of genome sequences. For example, the sequence of the *Mycoplasma genitalium* (500bk) genome was obtained with 96% coverage at 99.96% accuracy in a single run (Clarke, 2005). Such tools for interrogating the genetic blue-print of life on earth are of particular interest to microbial ecologists, as they can be applied to the genomic study of yet to be cultivated organisms. The wealth of output that they create, though, will mean we need to devise ever more powerful bioinformatic approaches.

The virtuous cycle: questions, data, informatics

The cycle of improved data acquisition followed by the invention of continually more sophisticated informatics solutions has characterized the genomic era from its beginnings in 1995 when the first bacterial genome was shotgun sequenced (Fleischmann *et al.*, 1995). Data acquisition, driven by the desire to ask ever more ambitious questions (often through a reduction in the cost of obtaining suitable data) is only the first step towards knowledge. Processing data has increasingly become the domain of

computers, and growing stocks of data mean we need equivalent advances in our ability to store, process, and integrate this data in biologically meaningful ways.

Data processing and analysis of environmental DNA (Tyson et al, 2004; Tringe et al, 2005) follows similar steps with the processing of shotgun sequences generated from isolate genomes (i.e. assembly of shotgun reads, gene and functional prediction), however is most often incomplete (fragmented) and from a mixed community, which makes it significantly harder than the processing and management of complete genomic datasets of laboratory isolates (Foerstner *et al.*, 2006). Therefore, ecologists and environmental biologists working in this area need improved methods of assembly and annotation to address the specific problems inherent to the metagenomic data types. Assembly of the environmental sequences is confounded by the lack of phylogenetic information associated to the scaffolds, contigs and shrapnel (unassembled reads) sequences. The grouping (binning) of anonymous sequenced fragments to known phylogenetic groups facilitates analysis by allowing the assignment of specific components of the predicted genes (and their functional capacity) to specific phylotypes. Additional problems in the analysis of the metagenomic data include the large numbers of hypothetical or orphan predicted proteins (proteins without known homologues), the need for vastly increased computational resources to process such sequence information (Edwards, 2006), and finally, the lack of easy-to-access descriptive data about genomes (Field *et al.*, et al, in review).

Towards next generation eco- informatic resources

Although the processing of metagenomic data a is associated with the numerous problems inherent to their incomplete nature, the currently existing methods of analysis can still provide important insights for the complexity and the functional capacity of the microbial ecosystems under study. The need for

new methods and informatic resources to facilitate eco-genomic studies is growing as the amount of data being generated from relevant microorganisms increases. It is within this backdrop that the projects discussed here are evolving.

The IMG analysis systems

As the genomic community is rapidly moving towards the generation of complete and draft sequences for several hundred genomes, it is becoming evident that the single most important tool for understanding the biology of a newly sequenced genome is the ability to effectively integrate it with available genome sequences to support comparative genomic data analysis. This approach follows the notion that it is in principle easier to annotate 1000 genomes than a single one as was originally proposed by Ross Overbeek (Overbeek, 2003).

The Joint Genome Institute's IMG system (http://img.jgi.doe.gov/) is a major effort to establish such a database (Markowitz *et al.*, 2006). The IMG was originally released on March 2005, and since then follows a quarterly release update schedule for both data and content. The current version of IMG (IMG 1.5, as on June 1, 2006) contains a total of 741 genomes consisting of 435 bacterial, 32 archaeal, 15 eukaryotic genomes and 259 bacterial phages. The IMG provides an integrated environment that facilitates the genomic analysis of the isolate organisms on a comparative level. The effectiveness of the comparative analysis depends on the availability of analytical tools and the efficiency of the integration. The latter in turn depends on the phylogenetic diversity of the organisms, the quality of the annotations and the level of detail in their cellular reconstructions. A key aspect of this system is the simplicity of the user interface which allows navigation among three major dimensions, *genomes*, *genes* and *functions*. In principle, all the genomic data in the IMG can be organized and compared using these three dimensions each of which further supports several associated data types. For example

the *genomes* dimension can provide access to IMG genomes organized based on Phylogenetic, Phenotypic or Ecotypic properties, etc.

More recently, an experimental system that stores metagenomic datasets, or 'microbiomes has also been launched by the JGI. The IMG/M system (Markowitz *et al.*, 2006b) was first released on March 2006, and essentially represents an expansion of IMG to include environmental data. The integration of metagenomic sequences in a data management system such as IMG, which was specifically designed for the analysis of isolate genomes, has revealed that such an approach is in fact valid, and analysis can be successfully performed. Further, this process has pointed out that such integration of genomic and metagenomic data can facilitate the identification of many problems and errors generated due to the above mentioned restrictions and shortcomings of the currently available methods, especially the incomplete nature of the sequences.

The metagenomic data in IMG/M can be also accessed through the three major dimensions. However, an additional level of complexity is introduced which is related to the heterogeneity of the phylogenetic classification of each metagenomic project. In this case, a single environmental project (metagenome or microbiome) does not correspond to a single phylogenetic point at the *genomes* dimension as it does in the IMG. Rather, the *genomes* dimension in the IMG/M becomes a separate three dimensional data space of its own, encompassing the canonical IMG's dimensions of *genomes, genes* and *functions*. It is within this "reduced" three dimensional data space that the curation which aims to dissect the environmental sample into its individual phylotypic components (i.e. binning) and their corresponding genetic and functional makeup, takes place. Yet, the ultimate goal of every environmental project lies in the delineation and understanding not only of its parts (i.e. the individual organisms) but also of the sum (i.e. the whole ecosystem). To this extent it is also important to employ tools and methods of comparative analysis across entire metagenomic projects as if they were individual organisms. This is

facilitated in the IMG/M by the “upper level” three dimensional data space, where the other two data dimensions of IMG (*genes* and *functions*) correspond to the total genetic and functional repertoire of the sequenced environmental sample.

Evidently, the availability of a data management system for isolate genomes, like IMG, was a prerequisite for the development of a corresponding metagenomics system. Indeed, the successful analysis and characterization of any environmental sample is directly dependent on phylogentic coverage in the sequenced isolate genome collection. In turn, the availability of a metagenome data management system is now offering valuable ideas and insights for the future development and data organization of analogous systems for isolate genomes. The large complexity of the individual organisms found in a single metagenomic project (either at the level of a strain or at any other phylogenetic level) can provide a window into the future of the isolate genome projects. Indeed, as the number of genome sequencing project increases exponentially, so does the need for databases that can efficiently organize and present them, either alone or in meaningful biological groupings (i.e. phylogenetic, phenotypic, ecotypic, etc.). Therefore, the efficient study and understanding of the organization and structure of the metagenomic data today is expected to play a major role in shaping the nature and structure of the genomic data and their organization in the near future.

The efficient capture of the genomic and environmental associated “metadata” (i.e. Phenotypes, Ecotypes, etc), is also expected to play a significant role in the future development of new methods of comparative analysis. As Microbial Ecology is becoming the center of an innovative energy at the interplay of Genomics and Ecology, we envision a stage where any attempt to understand the properties and biology of the individual organism would be intimately associated with its natural environment. At that point, the biologist’s dream of understanding the interconnections between genotype-environment-phenotype, should start to be more easily realized.

The key to achieving this goal is to have microbial ecologists, physiologists and biochemists, working together with bioinformaticians to create a comprehensive controlled vocabulary (a formalized set of terms) to describe the phenotypic properties of both single organisms as well as entire environmental communities. In parallel, intelligent computational systems should be designed to capture and disseminate these interactions and facilitate the comparative study of underlying patterns. The availability of associated metadata both for the isolate organisms as well as for whole environments, should then allow the design of new comparative analysis methods. Accordingly, a user could identify a specific profile within an environmental sample (defined as the combination of specific phenotypic properties within the sample), and then query this against a database of isolate genomes, to identify those that carry the combination of these properties.

The need for improved descriptions of genomes and metagenomes

As mentioned above, a key goal of the IMG is to place these genomes into proper organismal context. Information on the taxonomy, ecology, metabolism, and relevance also make it further possible to group, sort, and compare the features of these genomes in a variety of insightful ways. More importantly, those wishing to browse these collections and use them in large-scale comparative eco-genomic studies need rich contextual information, for example to explore the relationships between features of lifestyle and genomic content or structure. The IMG and IMG/M currently import such

data from the Genomes Online Database (GOLD) which has recently expanded its store of genomic information curated from the primary literature (Liolios *et al.*, 2006),

Developers of the IMG systems and GOLD, as well as a number of other database projects interested in this type of top-level information about genomes and metagenomes, have recently come together to form the Genomic Standards Consortium (GSC) (Field *et al.*, et al, in review). The GSC includes representatives from major genome sequencing centres, researchers generating and analyzing genomes and metagenomes, evolutionists, ecologists, database developers, computer scientists and bioinformaticians. Together this group is working towards the definition of a "Minimum Information about a Genome Sequence" (MIGS) specification. It is largely environmental and ecological information that is missing from the descriptions of our complete genome collections (Martiny and Field, 2005). Therefore, the involvement of ecologists who are knowledgeable about specific taxa and microbial communities and their interactions with the biotic and abiotic world is crucial. While the physical instantiation of the specification will emerge due to the collaboration of bioinformaticians, computer scientists, and experts in the development of ontologies and community standards, the content must be driven by biologists – and in particular ecologists.

The formation of the GSC was motivated by the growth in the number of genomes from environmental isolates and metagenomes. It is only through the collection of an expanded set of 'minimum information' about genomes and metagenomes that we will be able to readily collect information of particular use in eco-genomic studies. For example, the MIGS specification calls for information on the location, habitat, and biotic interactions of genomes to be recorded. It also calls for information like the primary citation for the isolation of the biological sample used. With exact information on location - namely latitude/longitude/altitude(depth) - we can place biological samples and their

molecules in a global reference. We can also combine this information with a vast array of geo-spatial information from a variety of sources (Lombardot *et al.*, 2006; Morrison *et al.*, 2006).

It is not only the description of genomic and metagenomic sequences that must be improved if we are to produce annotations that are useful for the proper re-interpretation, mining and integration of environmental experiments. For example, the environmental genomics community in the UK has recently proposed an extension of MIAME ("Minimum Information about a Microarray Experiment") (Brazma *et al.*, 2001), the accepted standard for the description of transcriptomic experiments. This extension, termed MIAME/Env, captures information on location, environmental conditions, biological treatments and phenotype of the samples under study (Morrison *et al.*, 2006). The GSC has drawn on the Env specification to develop the MIGS specification and the Metabolomics Standards Initiative (MSI) is evaluating Env for the sake of describing metabolomic datasets (http://msi-workgroups.sourceforge.net/).

Bioinformatics services, training and outreach

The emergence of new complex technologies, such as those seen in the domain of 'omics, brings with them the need for researchers to become familiar with a range of experimental and analytical tasks. In response to the need of researchers for specialized access to computing environments for bioinformatics, the formation of bioinformatics, service and research based centres of expertise, has become common place across scientific institutions. There are also a growing number of centres that also serve the specialist needs of a particular community.

One such centre providing support for the environmental genomics community is NEBC, a bioinformatics-based data centre established to promote the uptake of 'omic technologies in the environmental genomic community across the United Kingdom. NEBC works to support the data policy of the Natural Environment Research Council (NERC) which states that researchers must submit all NERC-funded data to public repositories or to NEBC in the event that no such repository exists (Tiwari *et al.*, 2006). To do this, NEBC provides bioinformatics support, in particular through a freely available computing platform called Bio-Linux which contains a large number of bioinformatics tools that can evolve according to the needs of the community (Field *et al.*). NEBC also works in the area of standards development and compliance (Morrison *et al.*, 2006), and participates in community-building activities, such as the organization of workshops that bring researchers and data management experts together (Field *et al.*, 2006b).

Conclusions

We are entering the next chapter of the genomic revolution in which the use of genomic and metagenomic approaches provide insights into the natural world on an unprecedented scale. This poses both unprecedented opportunities and challenges. The projects described here provide examples of how the ecological and environmental genomic communities can work together to exploit, adapt, and extend informatic resources to suit their own needs. The construction of suitable databases, richer standards for the reporting of experimental results, and adequate provision of bioinformatics training and support are all means to accelerate and improve research efforts in this area. It is clear that we

have as of yet merely glimpsed at what 'omic technology, judiciously applied and rigorously validated, will produce in the way of scientific discoveries. We predict that ecologists will have an increasingly important role to play in shaping the future of 'omic research in the next decade and beyond.

Acknowledgements

DF and NK thank the organizers of the MicroEnvGen II meeting (including BM and KN) for the opportunity to attend and present at this exciting workshop. NK is supported by the DOE's Office of Science, Biological and Environmental Research Program; the University of California, Lawrence Livermore National Laboratory, under contract no. W-7405-Eng-48; Lawrence Berkeley National Laboratory under contract no. DE-AC03-76SF00098; and Los Alamos National Laboratory under contract no. W-7405-ENG-36.

References

Brazma A, Hingamp P, Quackenbush J, *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29**, 365-371.

Clarke SC (2005) Pyrosequencing: nucleotide sequencing technology with bacterial genotyping applications. *Expert Rev Mol Diagn.* **5**, 947-953.

Edwards R (2006) Random Community Genomics. *http://phage.sdsu.edu/~rob/Edwards_RCG.pdf*.

Edwards RA, Rodriguez-Brito B, Wegley L, *et al.* (2006) Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions. *BMC Genomics.* **7**, 57.

Field D, Garrity GM, Morrison N, Selengut JD et al. (in review) Towards a richer description of genomes and metagenomes: creation of a "Minimal Information about a Genome Sequence" specification. *Nature Biotechnology*.

Field D, Tiwari B, Booth T, *et al.* (2006a) Open Source software for Biologists: from Famine to Feast. *Nature Biotechnology* **(in press)**.

Field D, Tiwari B, J. S (2005) Bioinformatics and Data Management support for Environmental Genomics. *PLoS Biol* **3**, e297.

Field D, Tiwari B, Snape J (2006b) Meeting Report: eGenomics: Genomes and the Environment. *Comparative and Functional Genomics* **6**, 363-368.

Fleischmann RD, Adams MD, White O, *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512.

Foerstner KU, von Mering C, Bork P (2006) Comparative analysis of environmental sequences: potential and challenges. *Philos Trans R Soc Lond B Biol Sci.* **361**, 519-523.

Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.* **34**, D332-334.

Lombardot T, Kottmann R, Pfeffer H, *et al.* (2006) Megx.net--database resources for marine ecological genomics. *Nucleic Acids Res.* **34**, D390-393.

Markowitz VM, Korzeniewski F, Palaniappan K, *et al.* (2006) The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34**, D344-348.

Markowitz VM, Ivanova N, Palaniappan K, *et al.* (2006) An experimental metagenome data management and analysis system. *Bioinformatics* (in press).

Martiny JBH, Field D (2005) Ecological Perspectives on our complete genome collection. *Ecology Letters.* **8**, 1334-1345.

Morrison N, Wood J, Hancock D, *et al.* (2006) Annotation of environmental 'omic data - Application to the transcriptomics domain. *OMICS: A Journal of Integrative Biology* **(in press)**.

Nelson KE, Paulsen IT, Heidelberg JF, Fraser CM (2000) Status of genome projects for nonpathogenic bacteria and archaea. *Nat Biotechnol* **18**, 1049-1054.

Overbeek R. (2003) The Project to annotate 1000 genomes, *http://www.thefig.info/archives/2004/01/manifesto_1.html*

Poinar HN, Schwarz C, Qi J, *et al.* (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science.* **311**, 392-394. Epub 2005 Dec 2020.

Shendure J, Mitra RD, Varma C, Church GM (2004) Advanced sequencing technologies: methods and goals. *Nat Rev Genet.* **5**, 335-344.

Shendure J, Porreca GJ, Reppas NB, *et al.* (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science.* **309**, 1728-1732. Epub 2005 Aug 1724.

Tiwari B, Field D, Snape J (2006) Public Data Repositories Need Serious Funding. *Nature* **439**, 912.

Tringe S, von Mering C, Kobayashi A, Salamov A, Chen K, et al. 2005. Comparative metagenomics of microbial communities. Science, 308, 554– 557.

Tyson, G.W., Chapman, J., Hugenholtz, P., et al. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature, 428, 37-43.

van Veen H, Kowalchuk G (2006) Microbial Environmental Genomics: Summary report of MicroEnGen-II (Shanghai, June 12-15, 2006). http://www.icsu-scope.org/projects/cluster2/microbial.htm.

Xu J (2006) Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Mol Ecol.* **15**, 1713-1731.

Zhang K, Martiny A, Reppas NB, *et al.* (2006) Sequencing genomes from single cells by polymerase cloning. *Nature Biotechnology* **24**, 680 - 686.